Zero-shot Learning via Discriminative Dual Semantic Auto-encoder

Xing, Nan; Liu, Yang; Zhu, Hong; Wang, Jing; Han, Jungong

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2020.3046573

Publication date:
2021

Citation for published version (APA):
Xing, N., Liu, Y., Zhu, H., Wang, J., & Han, J. (2021). Zero-shot Learning via Discriminative Dual Semantic Auto-encoder. *IEEE Access*, *9*, 733-742. [9303397]. https://doi.org/10.1109/ACCESS.2020.3046573

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 2, 2020, accepted December 17, 2020, date of publication December 22, 2020, date of current version January 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046573

Zero-Shot Learning via Discriminative Dual Semantic Auto-Encoder

NAN XING[1], YANG LIU[2], HONG ZHU[1], JING WANG[3], AND JUNGONG HAN[4]

[1]School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China
[2]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[3]Faculty of Printing Packaging Engineering and Digital Media Technology, Xi'an University of Technology, Xi'an 710048, China
[4]Computer Science Department, Aberystwyth University, Aberystwyth SY23 3FL, U.K.

Corresponding author: Yang Liu (yangl@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906141 and Grant 61705179, in part by the National Natural Science Foundation of Shaanxi Province under Grant 2020JQ-317, in part by the China Postdoctoral Science Foundation under Grant 2019M653564, in part by the Open Project Program of the State Key Lab of CAD&CG, Zhejiang University, under Grant A2018, in part by the Overseas Scholars' Science and Technology Activities Funding Project of Shaanxi Province, in part by the Fundamental Research Funds for the Central Universities, in part by the Doctoral Research Fund of Xi'an University of Technology under Grant 103-451119003, and in part by the Xi'an Science and Technology Foundation under Grant 2019217814GXRC014CG015-GXYD14.11.

ABSTRACT Zero-shot learning (ZSL) is an effective method to perform the recognition task without any training samples of specific classes. Most existing ZSL models put emphasis on learning an embedding between visual space and semantic space directly. However, few ZSL models research whether the human-designed semantic features are discriminative enough to recognize different classes. Moreover, one-way mapping suffers from the project domain shift problem. In this article, we propose to learn a Discriminative Dual Semantic Auto-encoder (DDSA) based on the encoder-decoder paradigm to solve this problem. DDSA attempts to construct two bidirectional embeddings to connect the visual space and the semantic space with the help of the learned aligned space which includes discriminative information of the visual features and semantic features. Based on the DDSA, we additionally propose a Deep DDSA to capture deep aligned features that are more conducive to zero-shot classification. The key to the proposed framework is that it implicitly exact the principal information from visual space and semantic space to construct aligned features, which is not only semantic-preserving but also discriminative. Extensive experiments on five benchmarks (SUN, CUB, AWA1, AWA2 and aPY) demonstrate the effectiveness of the proposed framework with state-of-the-art performance obtained on both conventional ZSL and generalized ZSL settings.

INDEX TERMS Zero-shot learning, discriminative, encoder-decoder, aligned.

I. INTRODUCTION

There are about 30,000 basic object categories and subordinate ones that human can recognize in the world. Human can even recognize new classes dynamically from few examples with little effort, but it is not easy for computer-based machine learning models that usually require thousands of labelled samples for training. It is well known that collecting enough training samples is time-consuming and labor-intensive. Motivated by the ability of humans to recognize unseen examples, the research area of zero-shot learning (ZSL) has received increasing interests, which aims to make good use of previously learned knowledge to recognize new categories without the need for labelled training data.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

Compared with the supervised learning, ZSL accomplishes the recognition task by using the semantic information [1] to build a relationship from seen classes to unseen classes. Moreover, test samples can also be considered from both seen and unseen categories, which is called Generalized Zero-Shot Learning (GZSL). In real-world applications, seen categories are usually more common than unseen ones, thus the GZSL is more realistic and challenging than ZSL for practical recognition tasks.

With respect to the bridge between the visual features and their corresponding semantics, most existing ZSL methods focus on building a mapping between visual space and semantic space with the seen samples. When classifying unseen samples, the mapping is used to project images of unseen class into the same semantic embedding space. Then the classifier is used in the embedding space to recognize new

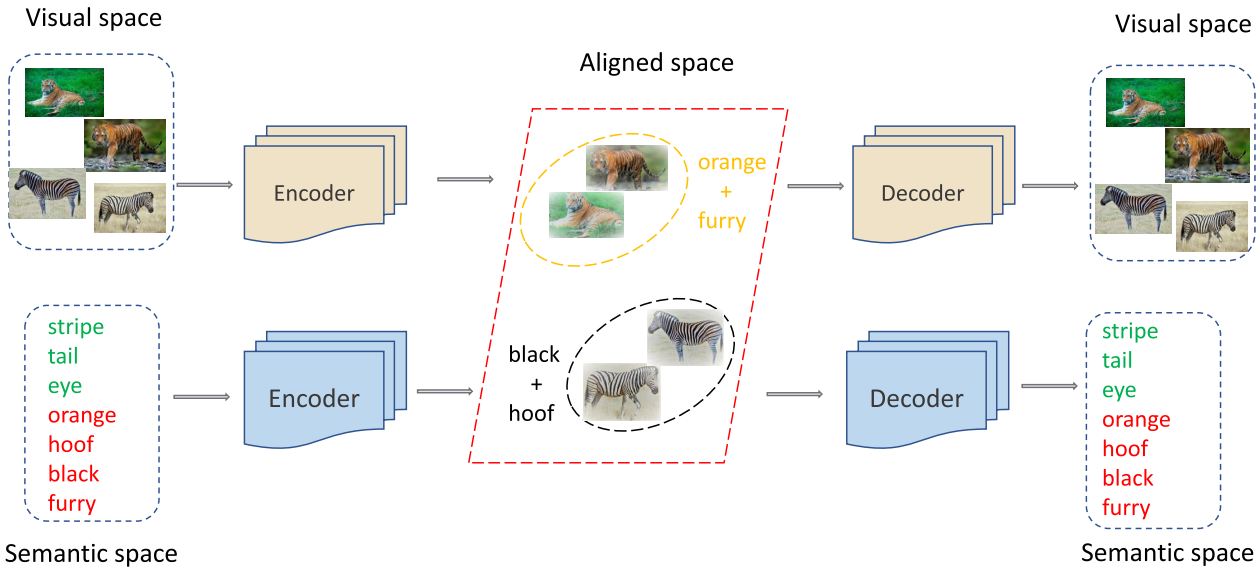


FIGURE 1. Illustration of the DDSA framework. Semantic attributes in green font are not discriminative, but attributes with red front have large variations.

samples from unseen classes, which is called the testing process. However, the unseen samples and the seen samples come from different classes, the mapping has been learned by seen samples easily generate a strong domain bias problem [2] when used in the unseen classes. This would lead to unseen samples being easily misclassified into seen classes. Moreover, most ZSL models ignore that whether the human-designed attributes are discriminative enough to recognize unseen classes. The large variations within each attribute will make it difficult to learn an appropriate classifier. Thus, the learned embeddings by these models cannot preserve the underlying discriminative information hidden in the seen classes.

In this article, we propose a framework named Discriminative Dual Semantic Auto-encoder (DDSA) to handle these problems. An example is shown in Figure 1 as the illustration of our framework. The framework intends to connect the three spaces *i.e.* visual space, aligned space and semantic space together by encoder-decoder paradigm. The learned aligned space can effectively remove irrelevant information (such as the background of *tiger* or *zebra* in Figure 1) in the visual space. Moreover, the aligned space can capture discriminative attribute correlations. For example, in Figure 1, the aligned space provides the possibility of finding the combinations of *orange+furry* and *black+hoof*, which is discriminative to recognize different classes. Our main contributions are summarized as follows:

- The proposed framework can find a aligned space where irrelevant information in the visual space can be removed and the the semantic information can be preserved, which is more constructive to establish a reconstruction relationship with the semantic space.The aligned space can preserve the semantic information.
- The seen class classifier and the cross reconstruction are utilized to make the aligned attributes discriminative enough to pull the data from the same class together and push those from different classes away from each other.
- Empirical results on five widely-used data sets show both DDSA and Deep DDSA outperforms existing ZSL models on five benchmarks and the convergence analysis also shows the stability of the proposed algorithm.

The rest of the article is organized is follows: Section 2 reviews the related work. Section 3 describes the proposed approaches DDSA and Deep DDSA in detail. Section 4 is reports experimental evaluation with some pertinent discussions. Finally, section 5 gives the conclusion.

II. RELATED WORKS

In this section, considering the key procedure in zero-shot learning, we mainly introduce the related works for ZSL from two aspects: the semantic information and visual-semantic embedding.

A. SEMANTIC INFORMATION

Unlike supervised learning, semantic information is a bridge to connect the seen and the unseen classes and it plays a key role in ZSL to make the recognition possible. The semantic information used in most works are attribute [3] and word-vector [4]. The attribute is characteristic descriptions of a class or an instance. For example, ''stripe'' can be shared between ''zebra'' and ''hyena'' and ''horselike'' can be shared between ''donkey'' and ''horse''. Thus attributes make it possible for ZSL to recognize novel classes in the world. Wordvector is a kind of semantic representation extracted directly from numerous text information such as Wikipedia articles and so on. Wordvector can also describe the differences and similarities between categories so it can build up

the relationship between seen and unseen classes. Recently, some deep models [5], [6] have been proposed to obtain more discriminative semantics by deep neural networks.

However, the collected human-designed semantic information is limited and redundant, thus the attribute or wordvector obtained are usually less discriminative to classify unseen classes. The limitation will create the domain gap among classes and result in a domain shift problem [2]. In this article, we propose to construct an aligned space to capture the discriminative information lying in the visual and semantic space.

B. VISUAL-SEMANTIC EMBEDDING

Semantic embedding aims to learn the mapping between the visual feature space and the semantic space with different semantic representations. According to different mapping directions, Visual-Semantic Embedding (VSE) framework can be divided into three types.

(1) *Visual→Semantic Embedding* aims to learn an embedding function from the visual feature space to the semantic space either using deep neural network ranking/regression [7]–[9] or via conventional ones [10]–[12]. For example, the DEVISE model [2] uses CNN and Word2Vec features as input to construct a deep zero-shot classification model. Reed *et al.* [13] trained an end-to-end framework to align with the fine-grained and category-specific content of images. Liu *et al.* [14] proposed a Graph and Auto-encoder based Feature Extraction (GAFE) model which brings the idea of auto-encoder into ZSL.

(2) *Semantic→Visual Embedding* aims to learn an embedding function from the semantic space to the visual feature space, such as [15], [16], which can effectively reduce the hubness problem. The training and testing process are similar with the *Visual→Semantic Embedding*. For example, DEM [17] projects the semantics of seen samples to the correspondding visual space by a deep embedding model to alleviate the hubness problem. Different from directly learning a mapping in visual space, Sung *et al.* [18] recently proposed to compare the visual features with embedded semantics with the help of a relation network (RN). It is noticed that RN tries to search for corresponding semantics in a self adaptive way.

(3) *Visual→Common Space←Semantic* aims to learn a common space where both the semantic space and the visual feature space are projected to, such as [19], [20]. In the testing phase, both attributes and visual features need to be embedded into the common space for classification. For example, Akata *et al.* [21] tried to learn a common joint space between semantics and visual features. Zhang *et al.* [22] learns a common embedding for visual and semantic features to get mixture patterns which are used to measure the similarity.

Different from such existing approaches which learn a single direction mapping between the visual space and the semantic space directly, we consider the encoder-decoder paradigm and try to learn two bidirectional mappings with the help of the constructed aligned space in this work.

III. APPROACH

A. PROBLEM DEFINITION

Given n labeled images with c seen classes $\{X, S, Y\}$ and n_u unlabeled images with c_u unseen classes $\{X_u, S_u, Y_u\}$. $X \in R^{d\times n}$ and $X_u \in R^{d\times n_u}$ represent d-dimensional seen and unseen visual features, while their corresponding labels are denoted by Y and Y_u, respectively. The labels of seen and unseen images do not have overlap, i.e., $Y \cap Y_u = \emptyset$. $S \in R^{k\times n}$ and $S_u \in R^{k\times n_u}$ are k-dimensional semantic representations of images in the seen and unseen datasets. For the zero-shot classification, our purpose is to learn a classifier $f : X_u \rightarrow Y_u$, where all unseen visual features X_u are completely unavailable during training.

B. DISCRIMINATIVE DUAL SEMANTIC AUTO-ENCODER

Traditional ZSL approaches mainly select the attribute space to perform the classification. However, there are two problems that should be considered. At first, the user-defined attributes are not always the same important for discrimination. Secondly, there are correlations among attributes, thus it is not suitable to use each attribute independently. To address such issue, in this work we build up the relationship between seen and unseen classes by learning dual auto-encoders: The first one *visual space↔aligned space* aims to learn an auto-encoder between visual space and aligned space. Meanwhile, the second one *aligned space↔semantic space* tries to learn an auto-encoder between aligned space and semantic space.

$L \in R^{m\times n}$ is used to represent the aligned space. In order to remove the irrelevant information in the visual space, we adopt a linear transformation $W \in R^{m\times d}$ to build up the relationship between aligned space and semantic space. Thus, the first auto-encoder between visual space and aligned space aims to solve the following function:

$$\min_{W,L} \|WX - L\|_F^2 + \left\|W^T L - X\right\|_F^2 \tag{1}$$

Moreover, to preserve the original semantic information, we adopt another linear mapping $Q \in R^{m\times k}$ to build up the relationship between aligned semantics and original semantics. Thus, the second auto-encoder can be formulated as follows:

$$\min_{L,Q} \|QS - L\|_F^2 + \left\|Q^T L - S\right\|_F^2 \tag{2}$$

In order to search discriminative attribute combinations to classify different classes, we adopt the a classifier of seen classes to make the learned aligned attributes more discriminative to zero-shot classification task. Specifically, an embeddding $P \in R^{c\times m}$ is learned from the aligned space to the label space. Finally, the objective of DDSA can be defined as follows:

$$\begin{aligned} \arg \min_{W,Q,P,L} & \|WX - L\|_F^2 + \left\|W^T L - X\right\|_F^2 \\ & +\alpha(\|QS - L\|_F^2 + \left\|Q^T L - S\right\|_F^2) + \beta \|PL - H\|_F^2 \\ s.t.\ & \|p_i\|_2^2 \leq 1, \quad \forall i, \end{aligned} \tag{3}$$

where $H = [h_1, h_2, \cdots, h_n] \in R^{c\times n}$ and $h_i = [0\cdots0, 1, 0\cdots0] \in R^c$ is a one-hot vector which represents the class label of the seen image x_i. P is a classifier in the aligned space. The last term in the objective tries to let the aligned attributes discriminative enough to classify different classes. In other words, it implicitly pull images from the same class together and pushes those from different classes away from each other.

1) OPTIMIZATION

Since Eq. (3) is not convex for W, Q, P and L simultaneously, it is difficult to the objective of DDSA directly. However, Eq. (3) is convex for each variable separately. We proposed to use an alternating optimization method to solve the objective of DDSA. Specifically, we alternate between the following subproblems:

Update W: We can update W by minimizing the following function:

$$W^* = \arg\min_W \|WX - L\|_F^2 + \left\|W^T L - X\right\|_F^2. \tag{4}$$

By taking the derivative of Eq. (4) and set it to zero, we can obtain:

$$\left(LL^T\right)W + W\left(XX^T\right) = 2LX^T. \tag{5}$$

We can see that the Eq. (5) is typically a Sylvester equation [23] that can be efficiently solved by a single line of code in MATLAB[1].

Update Q: Updating Q by minimizing the objective is equivalent to minimizing the following function:

$$Q^* = \arg\min_Q \|QS - L\|_F^2 + \left\|Q^T L - S\right\|_F^2. \tag{6}$$

Then Q can be optimized by solving the Sylvester function:

$$\left(LL^T\right)Q + Q\left(SS^T\right) = 2LS^T. \tag{7}$$

The solution of this problem is consistent with the Eq. (5).

Update P: We can update P by minimizing the following function:

$$\begin{aligned} &P^* = \arg\min_P \|PL - H\|_F^2 \\ &s.t. \quad \|p_i\|_2^2 \le 1, \quad \forall i. \end{aligned} \tag{8}$$

The Eq. (8) can be optimized by the Lagrange dual. Thus the analytical solution for Eq. (8) is:

$$P = \left(HL^T\right)\left(LL^T + \Lambda\right)^{-1}, \tag{9}$$

where Λ is a diagonal matrix constructed by all the Lagrange dual variables.

Update L: We can update L by minimizing the following function:

$$L^* = \arg\min_L \|A - BL\|_F^2, \tag{10}$$

[1] $W = sylvester(LL^T, XX^T, 2LX^T)$;

where

$$A = \begin{bmatrix} WX \\ X \\ \alpha QS \\ \alpha S \\ \beta H \end{bmatrix}, \quad B = \begin{bmatrix} I \\ W^T \\ \alpha I \\ \alpha Q^T \\ \beta P \end{bmatrix}, \tag{11}$$

and $I \in R^{m\times m}$ is the m-dimensional identity matrix. By taking the derivative of Eq. (10) and set it to zero, we can get the closed-form solution for L is:

$$L = \left(N^T N\right)^{-1} N^T M \tag{12}$$

In conclusion, The detailed procedure of solving the problem (3) is outlined in Algorithm 1. The optimization process always converges after tens of iterations in experiments and more details can refer to subsection 4.5.

2) VERIFICATION

In our experiment, we perform the classification task in visual space. At first, the semantic prototypes S_u are embedded into the visual space by the learned W and Q. Then the label of the testing image X_u^i can be classified by the Nearest Neighbour (NN) search with the help of following equation:

$$predict\ label\ (X_u^i) = \arg\min_j d\left(X_u^i, W^T Q S_u^j\right) \tag{13}$$

where X_u^i is the i-th sample of unseen images, and S_u^j is the semantic feature of the j-th unseen class. $d(\cdot,\cdot)$ represents the Euclidean distance between two vectors.

Algorithm 1 DDSA model for ZSL

Input:
Data matrix X, semantic matrix S, parameter α and β.
Initialization:
Q, P, L.
While not converged do:
1. Update W by solving Sylvester Eq. (5).
2. Update Q by solving Sylvester Eq. (7).
3. Update P by solving Eq. (9).
4. Update L by solving Eq. (12).
end.
Output:
W, Q, P, L.

C. DEEP DISCRIMINATIVE DUAL SEMANTIC AUTO-ENCODER

1) MULTIMODAL VARIATIONAL AUTOENCODER

In our work, VAE, an effective generative prototype, is employed as the basic building block of the proposed model. A standard VAE [24] is decomposed into an encoder that obtains low-dimensional latent variable z from input data x and a decoder that obtains output x' close to x from z. Typically, variational inference adopted in VAE aims to find

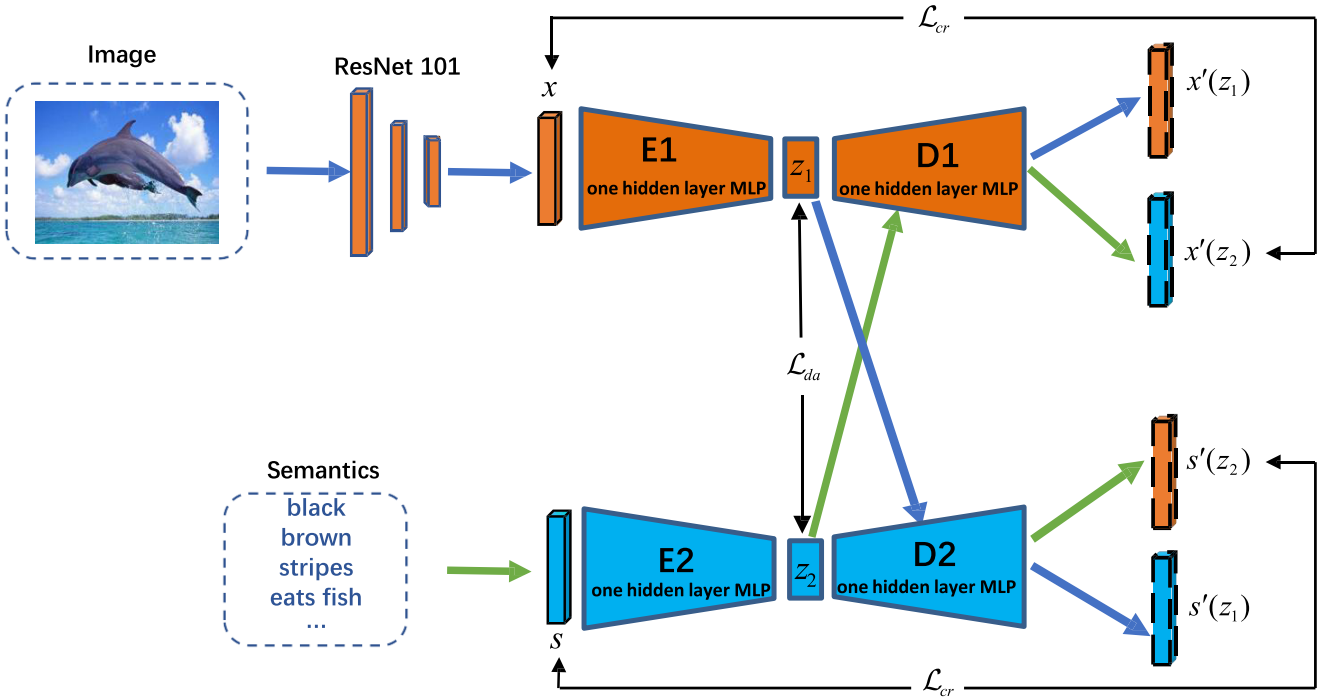


FIGURE 2. The framework of Deep Discriminative Dual Semantic Auto-encoder. Cross-reconstruction loss encourages the latent distributions to align ($\mathcal{L}_{cr}$). Moreover, distribution alignment in latent space is achieved by minimizing the Wasserstein distance between the latent distributions ($\mathcal{L}_{da}$).

the true conditional probability distribution $p_\theta(z|x)$ over the latent variable z. Due to the intractability of this distribution, its closest proxy posterior $q_\phi(z|x)$ acts as the approximation, through minimizing the distance of $q_\phi(z|x)$ and $p_\theta(z|x)$ using a variational lower bound limit. Thus, the objective function of a VAE is the variational lower bound on the marginal likelihood of a given data x, which can be formulated as:

$$\mathcal{L}_{vae} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\left(q_\phi(z|x)\|\, p_\theta(z)\right), \quad (14)$$

where the former term is the reconstruction error (REC) and the latter term is the Kullback-Leibler divergence (KL divergence) between $q_\phi(z|x)$ and $p_\theta(z)$. $p_\theta(z)$ is the prior distribution of z modeled as the multivariate Gaussian distribution. μ and $\sum$ are the mean and variance of the posterior distribution $q_\phi(z|x) = \mathcal{N}(\mu, \sum)$.

For our proposed method, a multimodal VAE (mVAE) structure is used to learn a shared latent embedding space of different modalities (visual features and semantic embeddings). As shown in Fig. 2, the encoder $E1$ and $E2$ transforms the visual feature $\tilde{x}$ and the semantic feature $\tilde{s}$ into low-dimensional latent vectors z_1 and z_2, respectively. Then, z_1 and z_2 are reconstructed into $x'(z_1)$ and $s'(z_2)$ by decoders D_1 and D_2, respectively. Formally, our mVAE sums the losses in the two modality-specific VAEs as:

$$\begin{aligned}\mathcal{L}_{mvae} &= \mathcal{L}^v_{vae} + \mathcal{L}^s_{vae} \\ &= \sum_{i=1}^{M} \mathbb{E}_{q_\varphi(z|x)}\left[\log p_\theta(x^{(i)}|z)\right] \\ &\quad -D_{KL}\left(q_\phi(z|x^{(i)})\middle\|\, p_\theta(z)\right), \end{aligned} \quad (15)$$

where $\mathcal{L}^v_{vae}$ and $\mathcal{L}^s_{vae}$ represent the VAE losses of visual and semantic modality, respectively. $M = 2$ denotes the two modality data, *i.e.,* visual feature and semantic feature.

2) CROSS-RECONSTRUCTION (CR) WITH LATENT EMBEDDINGS

In order to make the modality-specific autoencoder further learn similar representations across modalities, the proposed mVAE allow reconstructing the modality data of an instance by decoding the latent embeddings of a different instance from another modality of the same class. Intuitively, although the latent embeddings of the same classes come from different modalities, they should be semantically consistent. Thus, the cross-reconstruction loss of i-th visual feature x_i can be derived the following cross-modal softmax function:

$$\mathcal{L}^v_{cr}(x_i) = \frac{\exp(x_i^T D_1(E_2(s_i)))}{\sum_{j=1}^{c}\exp(x_i^T D_1(E_2(s_j)))} \quad (16)$$

where E_2 is the encoder of semantic features and D_1 is the decoder of visual features.

Similarly, he cross-reconstruction loss of i-th semantic feature s_i can be derived the following cross-modal softmax function:

$$\mathcal{L}^s_{cr}(s_i) = \frac{\exp(s_i^T D_2(E_1(x_i)))}{\sum_{j=1}^{n}\exp(s_i^T D_2(E_1(x_j)))} \quad (17)$$

where E_1 is the encoder of visual features and D_2 is the decoder of semantic features.

Our goal is to maximize the above probabilities in both the visual and semantic spaces, which can be formulated by

TABLE 1. Details of Datasets, Where s/u Means Seen/Unseen.

Dataset	visual dim	s/u classes	s/u samples
SUN	2000	645/72	10320/1440
CUB	2000	150/50	7057/2967
AWA1	2000	40/10	19832/5685
AWA2	2000	40/10	23527/7913
aPY	2000	20/12	5932/7924

minimizing the following multi-modal cross-entropy loss:

$$\mathcal{L}_{cr} = -\sum_{i=1}^{n} \mathcal{L}_{cr}^{v}(x_i) - \sum_{i=1}^{c} \mathcal{L}_{cr}^{s}(s_i) \tag{18}$$

3) DISTRIBUTION-ALIGNMENT (DA) IN LATENT SPACE

Generated images should match with the semantic features by minimizing their distance. In this model, 2-Wasserstein distance [25] is used as the alignment criterion between the latent Gaussian distributions of visual features and semantic features. Thus, the distance can be represented by:

$$W_{ij} = \Big[\left\| \mu_i - \mu_j \right\|_2^2 + tr(\sum\nolimits_i) + tr(\sum\nolimits_j) - 2(\sum\nolimits_i^{\frac{1}{2}} \sum\nolimits_i \sum\nolimits_j^{\frac{1}{2}})^{\frac{1}{2}} \Big]^{\frac{1}{2}}, \tag{19}$$

where i and j represent different features. As the diagonal covariance matrices predicted by an encoder is commutative, we further rewrite Eq. (19) in the following form:

$$W_{ij} = (||\mu_i - \mu_j||_2^2 + || \sum\nolimits_i^{\frac{1}{2}} - \sum\nolimits_j^{\frac{1}{2}} ||_F^2)^{\frac{1}{2}} \tag{20}$$

where $\|\cdot\|_F$ is the Frobenius norm. Finally, the distribution alignment loss function is derived as:

$$\mathcal{L}_{da} = \sum_{i}^{M} \sum_{j \neq i}^{M} W_{ij} \tag{21}$$

4) OVERALL OBJECTIVE FUNCTION

By taking the introduced above into account, the overall objective function of the Deep DDSA is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{mvae} + \gamma_1 \mathcal{L}_{cr} + \gamma_2 \mathcal{L}_{da}, \tag{22}$$

where γ_1 and γ_2 are trade-off parameters chosen based on the validation dataset.

IV. EXPERIMENTS

In this section, We validate our proposed specific linear and deep methods on five widely-used data sets and compared with some state-of-the-art models.

A. DATASETS DESCRIPTIONS

SUN Attribute (SUN) [26] consists of 14,340 images to describe 717 scene classes where 645 classes are selected as seen samples and the remaining 72 classes are unseen samples. For each class, a 102-dimension continuous attribute vector is provided.

Caltech-UCSD Birds-200-2011 (CUB) [27] contains 11,788 images of 200 fine-grained bird classes. A standard split divides these bird species into 150 classes for seen dataset and 50 for unseen dataset. A 312-dimension attribute vector is used for each class as semantic description.

Animals with Attributes 1 (AWA1) [3] contains 30,745 images of 50 classes of animals where 40 classes are selected as seen samples and the remaining 10 classes are unseen samples. A 85-dimension continuous attribute vector is used for each class as semantic description.

Animals with Attributes 2 (AWA2) [28] consists of 37,322 visual features and 85 class-level attributes. Similarly, 40/10 classes are selected for seen samples/unseen samples and all of the 50 categories are the same as AWA1 data set.

A Pascal and Yahoo (aPY) [29] is a small-scale coarse-grained data set with 64 attributes. It contains 32 classes, where 20 Pascal classes and 12 Yahoo classes are used for training and testing, respectively.

For all datasets, we follow the settings in [28] to split each dataset for training and testing. Moreover, for fair comparison, the visual feature of each sample is represented by 2048-dim vector extracted by 101-layered ResNet [30]. The statistics of all five datasets are given in Table 1.

B. IMPLEMENTATION DETAILS

For DDSA, parameters α and β in our objective function are fine-tuned in the range $[10^{(-3)}, 10^3]$ using the validation splits. Finally, we set the dimension of the aligned space is 1200, *i.e.* $m = 1200$. More details about the parameters can be seen in last subsection.

For Deep DDSA, the encoders (E_1 and E_2) and decoders (D_1 and D_2) are all implemented as multilayer perceptron (MLP) with only one hidden layer. In detail, 1560 and 1450 hidden units are used for the encoder E_1 and E_2, respectively. Meanwhile, 1660 and 660 hidden units are used for D_1 and D_2, respectively. Moreover, the latent embedding size is set to 90 in all datasets. We train the model for 100 epochs by the Adam optimizer and a batch size of 50 for all datasets. After training, the discriminative visual features and semantics from the seen samples and unseen samples are transformed into a shared latent space where the training and test set of the final classifier is performed. Finally, the proposed model is implemented in the deep learning toolkit "TensorFlow 1.3.0".

C. EVALUATION METRICS AND COMPARISON METHODS

The average per-class Top-1 accuracy is used for the evaluation criteria, which is formulated by

$$acc\,(\Upsilon) = \frac{1}{\|\Upsilon\|} \sum_{c=1}^{\|\Upsilon\|} \frac{\#correct\ predictions\ in\ c}{\#samples\ in\ c} \tag{23}$$

TABLE 2. ZSL Results on SUN, CUB, AWA1, AWA2 and aPY Datasets. The Results Report Average Per-Class Top-1 Accuracy in %.

Type	Method	SUN	CUB	AWA1	AWA2	aPY
Shallow	DAP	39.9	40.0	44.1	46.1	33.8
	IAP	19.4	24.0	35.9	35.9	36.6
	SSE	51.5	43.9	60.1	61.0	34.0
	LATEM	55.3	49.3	55.1	55.8	35.2
	SJE	53.7	53.9	65.6	61.9	32.9
	ESZSL	54.5	53.9	58.2	58.6	38.3
	SYNC	56.3	**55.6**	54.0	46.6	23.9
	SAE	40.3	33.3	53.0	54.1	8.3
	LESAE	60.0	53.9	66.1	68.4	40.8
	GAFE	62.2	52.6	67.9	67.4	44.3
	DDSA	**63.3**	53.2	**68.3**	**69.1**	**46.1**
Deep	DEVISE	56.5	52.0	54.2	59.7	39.8
	CONSE	38.8	34.3	45.6	44.5	26.9
	CMT	39.9	34.6	39.5	37.9	28.0
	SP-AEN	59.2	55.4	-	58.5	24.1
	PSR	61.4	56.0	-	63.8	38.4
	DCN	61.8	56.2	65.2	-	43.6
	CCSS	56.8	44.1	56.3	63.7	35.5
	f-CLSWGAN	58.5	57.7	64.1	-	-
	cycle-CLSWGAN	60.0	58.4	66.3	-	38.6
	CADA-VAE	61.8	60.4	62.3	64.0	35.7
	CCGN	62.8	53.5	70.6	71.6	43.8
	MAAE	62.1	-	71.4	69.4	43.2
	Deep DDSA	**64.6**	**61.5**	**71.5**	**72.0**	**44.8**

where Υ and $\|\Upsilon\|$ are defined as the set of categories and number of categories respectively. In other words, Υ consists of all the unseen classes, i.e. the testing classes.

The generalized zero-shot learning (GZSL) is another evaluation criteria, whose search space at testing time is not restricted to only testing categories (Υ^{ts}), but consists of the training ones (Υ^{tr}). In this case, we can compute $acc\left(\Upsilon^{ts}\right)$ and $acc\left(\Upsilon^{tr}\right)$ by Eq. (23). In addition, the harmonic mean can be computed as follows

$$H = \frac{2 \cdot acc\left(\Upsilon^{ts}\right) \cdot acc\left(\Upsilon^{tr}\right)}{acc\left(\Upsilon^{ts}\right) + acc\left(\Upsilon^{tr}\right)} \tag{24}$$

In the experiment, we compare the proposed model with many competitive or representative methods, including **shallow methods**: DAP [3], IAP [3], SSE [22], SJE [21], ESZSL [31], LatEm [32], SYNC [33], SAE [34], LESAE [35], GAFE [14] and some **deep methods**: DEVISE [2], CMT [4], CONSE [36], SP-AEN [11], PSR [7], DCN [37], CCSS [38], f-CLSWGAN [39], cycle-CLSWGAN [40], CADA-VAE [41], CCGN [42], MAAE [43].

D. EFFECTIVENESS OF THE PROPOSED FRAMEWORK

In order to demonstrate the effectiveness of each component in the objective function, we compare four different approaches and give the ZSL results on SUN dataset in Figure 3. (1) Only learn one auto-encoder between visual space and the aligned space with the help of discriminative constraint (**VD**) (*i.e.* 1, 2, 5 term in Eq. (3)); (2) Only learn one auto-encoder between semantic space and the aligned space with the help of discriminative constraint (**SD**) (*i.e.* 3, 4, 5 term in Eq. (3)); (3) Learn two auto-encoders between visual/semantic space and the aligned space but without discriminative constraint (**VS**) (*i.e.* 1, 2, 3, 4 term in Eq. (3)); (3) Learn two auto-encoders between visual/semantic space and

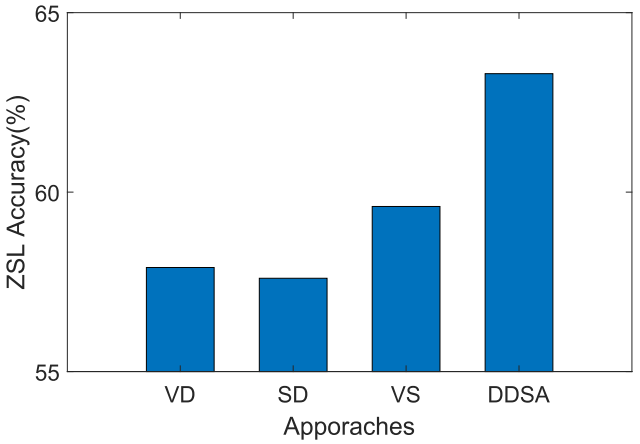


FIGURE 3. Comparisons of four approaches on SUN data set.

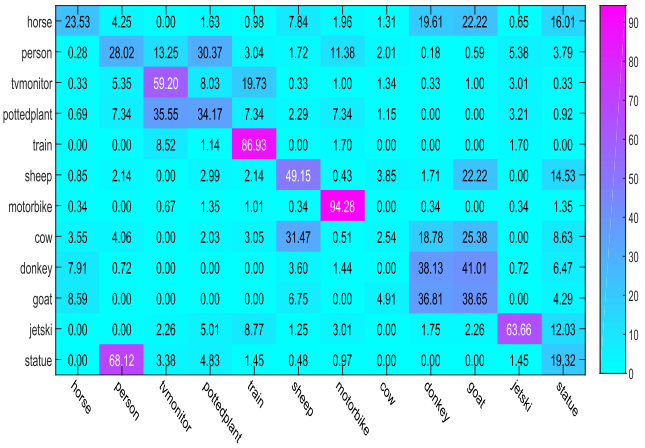


FIGURE 4. Confusion matrices of unseen classes for Deep DDSA on the aPY data sets. The Top-1 accuracy is between 0 and 100 (%).

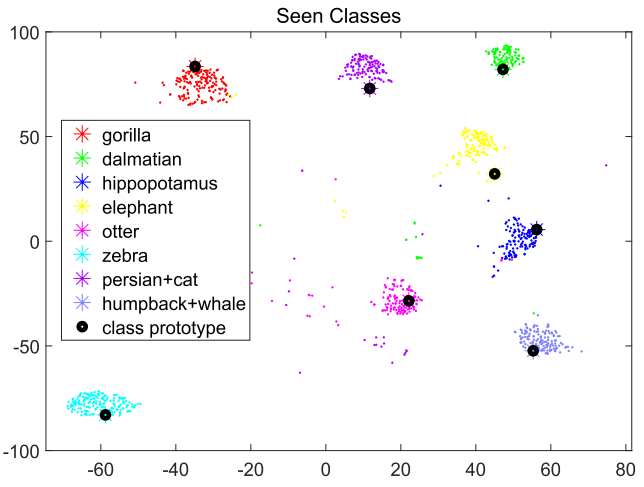


FIGURE 5. Visualization of prototypes and projected samples for Deep DDSA on the AwA2 data set in the semantic space by t-SNE.

TABLE 3. GZSL Results on SUN, CUB, AWA1, AWA2 and aPY Datasets. ts = Top-1 Accuracy of the Test Unseen-Class Samples, tr = Top-1 Accuracy of the Test Seen-Class Samples, H = Harmonic Mean (CMT*: CMT With Novelty Detection). We Measure Top-1 Accuracy in %.

		SUN			CUB			AWA1			AWA2			aPY		
Type	Method	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
Shallow	DAP	4.2	25.1	7.2	1.7	67.9	3.3	0.0	**88.7**	0.0	0.0	84.7	0.0	4.8	**78.3**	9.0
	IAP	1.0	37.8	1.8	0.2	**72.8**	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
	SSE	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
	LATEM	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
	SJE	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
	ESZSL	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
	SYNC	7.9	**43.3**	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	**90.5**	18.0	7.4	66.3	13.3
	SAE	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	0.4	80.9	0.9
	LESAE	21.9	34.7	26.9	24.3	53.0	33.3	19.1	70.2	30.0	21.8	70.6	33.3	12.7	56.1	20.1
	GAFE	19.6	31.9	24.3	22.5	52.1	31.4	25.5	76.6	38.2	26.8	78.3	40.0	15.8	68.1	25.7
	DDSA	**22.3**	33.9	**26.9**	**25.1**	53.9	**34.3**	**26.3**	77.1	**39.2**	**28.7**	82.8	**42.6**	**20.4**	62.1	**30.7**
Deep	DEVISE	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
	CMT	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
	CMT*	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	8.7	89.0	15.9	10.9	74.2	19.0
	CONSE	6.8	**39.9**	11.6	1.6	**72.2**	3.1	0.4	**88.6**	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
	PSR	20.8	37.2	26.7	24.6	54.3	33.9	-	-	-	20.7	73.8	32.3	13.5	51.4	21.4
	f-CLSWGAN	42.6	36.6	39.4	43.7	57.7	49.7	57.9	61.4	59.6	-	-	-	-	-	-
	cycle-CLSWGAN	49.4	33.6	40.0	45.7	61.0	52.3	56.9	64.0	60.2	-	-	-	-	-	-
	CADA-VAE	47.2	35.7	40.6	51.6	53.5	52.4	57.3	72.8	64.1	55.8	75.0	63.9	-	-	-
	CCGN	32.9	34.8	33.9	29.1	49.2	36.6	53.9	68.0	60.2	51.3	73.2	60.4	29.6	69.7	41.5
	MAAE	23.1	36.7	28.4	-	-	-	51.0	84.3	63.4	51.4	85.6	64.2	15.4	74.1	25.5
	Deep DDSA	**49.6**	36.1	**41.8**	**52.0**	53.9	**53.0**	**59.1**	71.9	**64.9**	**57.2**	76.3	**65.4**	**33.3**	67.3	**44.6**

the aligned space with discriminative constraint (**DDSA**) (*i.e.* Eq. (3)).

Figure 3 gives the performance of different strategies. By comparing the ZSL results of **VD**, **SD** and **DDSA**, we conclude that using dual auto-encoders is successful for the ZSL task. Moreover, by comparing the performance of **VS** and **DDSA**, we can see that imposing the discriminative constraint in the objective function can also improve the recognition accuracy.

E. ZSL AND GZSL RESULTS

Table 2 gives ZSL results of different methods. For DDSA, it achieves the best results on all datasets except the CUB dataset. Especially on the aPY dataset, the accuracy of DDSA increase 1.8% compared the strongest shallow competitor GAFE [14]. On the other three datasets (SUN, AWA1 and AWA2), the advantage of the DDSA is also obvious. The reason why DDSA does not achieve the highest recognition rate on the CUB database is that CUB is a fine-grained dataset where most classes are very similar, so less discriminative structure could be obtained by the DDSA. For deep DDSA, it consistently performs better than compared deep ZSL models on all datasets. Especially on the SUN dataset, the accuracies increase of 1.8% compared to the strongest deep competitor CCGN [42]. The promising performance of both DDSA and Deep DDSA suggests the the classification performance of unseen classes can be improved with the help of discriminative aligned attributes.

The GZSL results on five small-scale attribute datasets is shown in Table 3. According to the GZSL results, we get following observations:

(1) Compared with the ZSL results in Table 2, the GZSL classification accuracy (“ts” value) are lower than ZSL results. The reason is that all of the seen samples are included in the search space as interferences of test images.

(2) Low accuracy on “ts” value but high accuracy on “tr” implies some ZSL models such as DAP [3] and SYNC [33] perform well on seen classes but fails to generalize for novel (unseen) classes. On the other hand, the classification accuracy of most ZSL models on seen classes is higher than the accuracy on unseen classes, *i.e.*, $tr > ts$. The reason is that although these ZSL models are trained with the help of visual features of seen samples, but the predictability for unseen samples is still very poor.

(3) Table 3 shows that both DDSA and Deep DDSA achieve best results on the “ts” value and “H” value on all five datasets. Specifically, for “H” value, DDSA obtains 42.6% on AWA2 dataset and 30.7% on aPY dataset, which is better than the next best shallow model GAFE by 2.6% and 5.0%, respectively. In addition, compared with the closest baseline CADA-VAE [41], the accuracy difference is as follows: 41.8% vs 40.6% on SUN, 53.0% vs 52.4% on CUB, 64.9% vs 64.1% on AWA1, 65.4% vs 63.9% on AWA2. This demonstrates the superiority of Deep DDSA for the GZSL task.

F. VISUALIZATION OF CLASS STRUCTURE

We provide the visualized results and confusion matrix to make the result more understandable. The column and the row of the confusion matrix respectively represents the ground truth and the predicted results. According to Figure 4, it is clear that our Deep DDSA algorithm can identify most of unseen categories, except “horse”, “person”, “cow” and “statu” on the AWA2 data set.

t-SNE [44] is used to project samples and prototypes from semantic space to a 2-D plane. Its function is to display the distance between samples and the corresponding class

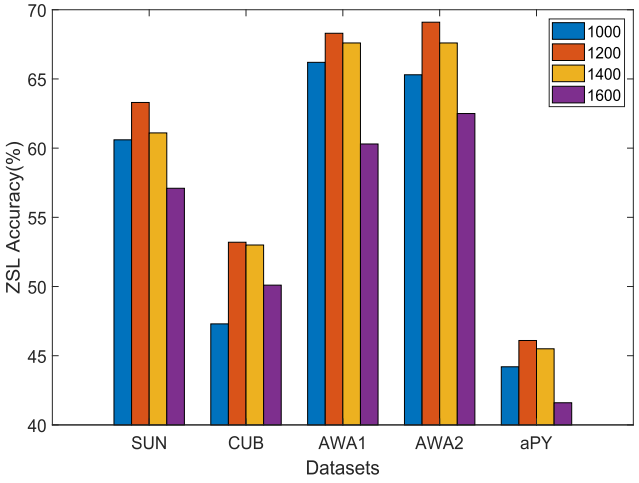


FIGURE 6. ZSL result of DDSA under different dimensions of aligned space on four data sets.

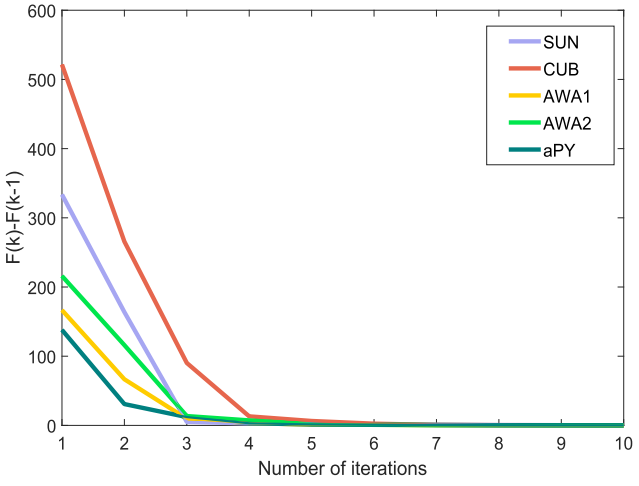


FIGURE 7. Convergence curve of DDSA on four datasets.

prototypes. Figure 5 shows that most samples locate near the prototypes of the corresponding classes, which demonstrates the proposed Deep DDSA algorithm can learn a proper projection from the visual feature space to the semantic space.

G. DIMENSION SETTING AND CONVERGENCE ANALYSIS

According to Figure 6, the selection of the aligned space'dimension has an influence on the recognition rate. Furthermore, the proposed model performs better when the dimension of the aligned space is 1200. In Figure 7, the $F(k)$ represents the F-norm of the objective function Eq. (3) after k-th iteration by Algorithm 1. It is easy to see the proposed linear model converges within only 7 steps on all data sets. On the other hand, the complexity of Eq. (5) and Eq. (7) depend on the dimension of aligned attributes *i.e.* $O\left(m^3\right)$ instead of the number of samples. Thanks to the low complexity and good convergence, the proposed model has a better practical application than most deep and shallow ZSL algorithms.

V. CONCLUSION

A novel ZSL framework named Discriminative Dual Semantic Auto-encoder (DDSA) is proposed in this work. This framework aims to learn an aligned attribute space where the irrelevant information hidden in the visual space can be removed and the semantic information can be preserved. Moreover, we proposed a Deep DDSA in order to capture deep features in the aligned attribute space. Empirical results on five widely-used data sets show both DDSA and Deep DDSA outperforms existing ZSL models on five benchmarks and the convergence analysis also shows the stability of the proposed algorithm.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and AE for their constructive comments and suggestions, which improve the quality of the article.

REFERENCES

[1] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 951–958.

[2] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[3] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[4] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 935–943.

[5] Z.-L. Yang, X.-Q. Guo, Z.-M. Chen, Y.-F. Huang, and Y.-J. Zhang, "RNN-stega: Linguistic steganography based on recurrent neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1280–1295, May 2019.

[6] W. Cai and Z. Wei, "Piigan: Generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2020.

[7] S. Biswas and Y. Annadani, "Preserving semantic relations for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7603–7612.

[8] Y. Liu, X. Gao, Q. Gao, J. Han, and L. Shao, "Label-activating framework for zero-shot learning," *Neural Netw., Off. J. Int. Neural Netw. Soc.*, vol. 121, pp. 1–9, Jan. 2020.

[9] J. Li, X. Lan, Y. Liu, L. Wang, and N. Zheng, "Compressing unknown images with product quantizer for efficient zero-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5458–5467.

[10] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, early access, Oct. 1, 2020, doi: 10.1109/LGRS.2020.3026587.

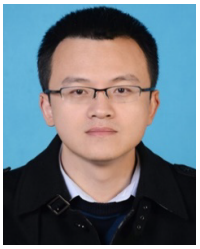
[11] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1043–1052.

[12] Y. Liu, J. Li, and X. Gao, "A simple discriminative dual semantic auto-encoder for zero-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 4053–4057.

[13] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[14] Y. Liu, D. Xie, Q. Gao, J. Han, S. Wang, and X. Gao, "Graph and autoencoder based feature extraction for zero-shot learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3038–3044.

[15] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2452–2460.
[16] H. Zhang and P. Koniusz, "Zero-shot kernel learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7670–7679.
[17] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3010–3019.
[18] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.
[19] Q. Wang and K. Chen, "Zero-shot visual recognition via bidirectional latent embedding," *Int. J. Comput. Vis.*, vol. 124, no. 3, pp. 356–383, Sep. 2017.
[20] V. K. Verma and P. Rai, "A simple exponential family framework for zero-shot learning," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2017, pp. 792–808.
[21] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2927–2936.
[22] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4166–4174.
[23] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation AX+ XB= C [F4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.
[24] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *CoRR*, vol. abs/1312.6114, pp. 1–14, Apr. 2014.
[25] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *Michigan Math. J.*, vol. 31, no.2, pp. 231–240, 1984.
[26] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, nos. 1–2, pp. 59–81, May 2014.
[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2011.
[28] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.
[29] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1778–1785.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.
[32] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 69–77.
[33] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5327–5336.
[34] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3174–3183.
[35] Y. Liu, Q. Gao, J. Li, J. Han, and L. Shao, "Zero shot learning via low-rank embedded semantic AutoEncoder," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2490–2496.
[36] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," 2013, *arXiv:1312.5650*. [Online]. Available: http://arxiv.org/abs/1312.5650
[37] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized learning with deep calibration network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2009–2019.
[38] J. Liu, X. Li, and G. Yang, "Cross-class sample synthesis for zero-shot learning," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 113.
[39] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.
[40] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. ECCV*, 2018, pp. 21–37.
[41] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8239–8247.
[42] L. Sun, J. Song, Y. Wang, and B. Li, "Cooperative coupled generative networks for generalized zero-shot learning," *IEEE Access*, vol. 8, pp. 119287–119299, 2020.
[43] Z. Ji, G. Dai, and Y. Yu, "Multi-modality adversarial auto-encoder for zero-shot learning," *IEEE Access*, vol. 8, pp. 9287–9295, 2020.
[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.



NAN XING received the Ph.D. degree from Xidian University, Xi'an, China, in 2018. He is currently an Associate Professor with the School of Automation and Information Engineering, Xi'an University of Technology. His current research interests include image processing and pattern recognition.

YANG LIU received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2013, 2015, and 2018, respectively. He is currently a Postdoctoral Researcher with Xidian University. He has authored nearly 20 technical articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON IMAGE, the IEEE TRANSACTIONS ON CYBERNETICS, PR, CVPR, AAAI, and IJCAI. His research interests include dimensionality reduction, pattern recognition, and deep learning.

HONG ZHU received the Ph.D. degree from Fukui University, Fukui, Japan, in 1999. She is currently a Professor with the School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China. Her research interests include image processing and pattern recognition.

JING WANG received the Ph.D. degree in pattern recognition and intelligent system from the Xi'an University of Technology, Xi'an, China. She is currently a Teacher with the Faculty of Printing Packaging Engineering and Digital Media Technology, Xi'an University of Technology. Her research interests include computer vision, image processing, and pattern recognition.

JUNGONG HAN is currently a Professor with the Computer Science Department, Aberystwyth University, U.K. His current research interests include multimedia content identification, multisensor data fusion, computer vision, and multimedia security. Dr. Han is an Associate Editor of *Neurocomputing* (Elsevier), and an Editorial Board Member of *Multimedia Tools and Applications* (Springer).